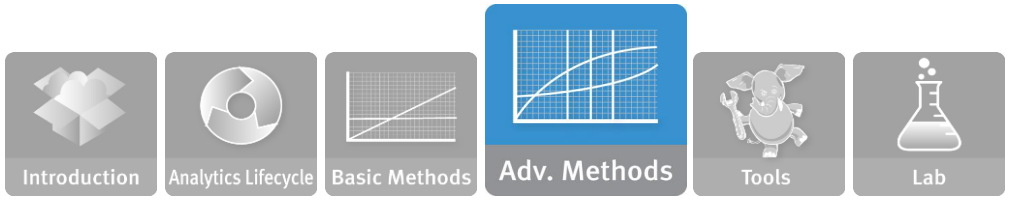
Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Module 4 – Advanced Analytics - Theory and Methods



Module 4: Advanced Analytics – Theory and Methods

Upon completion of this module, you should be able to:

- Examine analytic needs and select an appropriate technique based on **business objectives; initial hypotheses; and the data's structure and volume**
- Apply some of the more commonly used methods in Analytics solutions
- Explain the algorithms and the technical foundations for the commonly used methods
- Explain the environment (use case) in which each technique can provide the most value
- Use appropriate diagnostic methods to validate the models created
- Use R and in-database analytical functions to fit, score and evaluate models

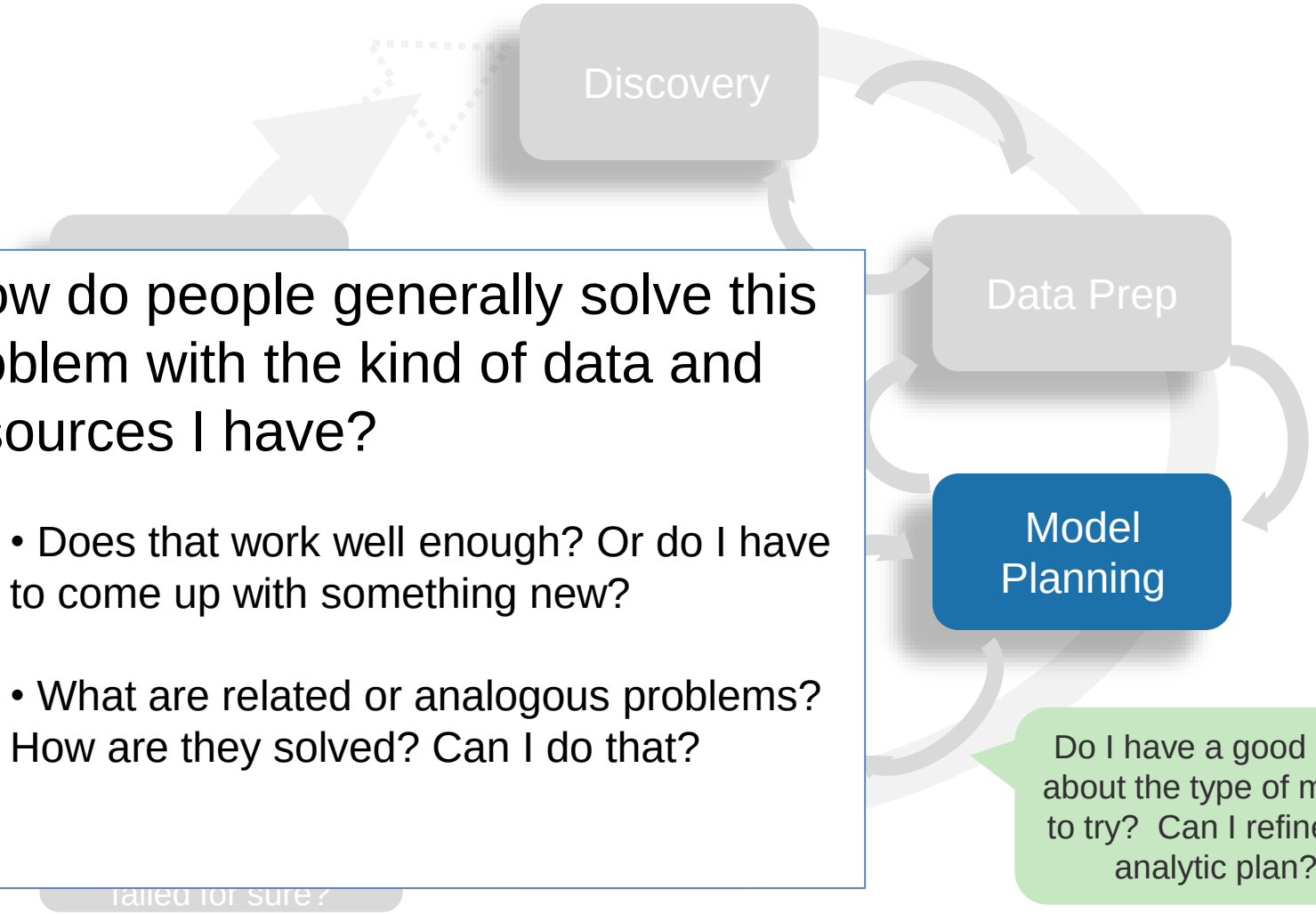
Where are we?

- In Module 3 we reviewed R skills and basic statistics
- You can use R to:
  - Generate summary statistics to investigate a data set
  - Visualize Data
  - Perform statistical tests to analyze data and evaluate models
- Now that you have data, and you can see it, you need to plan the analytic model and determine the analytic method to be used

Applying the Data Analytics Lifecycle

- In a typical Data Analytics Problem - you would have gone through:
  - Phase 1 – Discovery - have the problem framed
  - Phase 2 – Data Preparation - have the data prepared
- Now you need to plan the model and determine the method to be used.

Phase 3 - Model Planning



How do people generally solve this problem with the kind of data and resources I have?

- Does that work well enough? Or do I have to come up with something new?
- What are related or analogous problems? How are they solved? Can I do that?

What Kind of Problem do I Need to Solve? How do I Solve it?

The Problem to Solve	The Category of Techniques	Covered in this Course
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	Apriori
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	Naïve Bayes Decision Trees
I want to find the structure in a temporal process…predict output over time. I want to forecast the behavior of a temporal process	Time Series Analysis	ACF, PACF, ARIMA
I want to analyze my text data	Text Analysis	Regular expressions, Document representation (Bag of Words), TF-IDF

Why These Example Techniques?

- Most popular, frequently used:
  - Provide the foundation for Data Science skills on which to build
- Relatively easy for new Data Scientists to understand & comprehend
- Applicable to a broad range of problems in several verticals

Module 4: Advanced Analytics – Theory and Methods

part 1: K-means Clustering

During this lesson the following topics are covered:

- Clustering – Unsupervised learning method
- K-means clustering:
  - Use cases
  - The algorithm
  - Determining the optimum value for K
  - Diagnostics to evaluate the effectiveness of the method
  - Reasons to Choose (+) and Cautions (-) of the method

Clustering

The problem of **finding a hidden structure within unlabeled data.**

How do I group these documents by topic?

How do I group my customers by purchase patterns?

- Sort items into groups by similarity:
  - Items in a cluster are more similar to each other than they are to items in other clusters.
  - Need to detail the properties that characterize “similarity”
    - Or of distance, the "inverse" of similarity
- Not a predictive method; finds similarities, relationships
- **In a cluster we end up with a tight group (homogeneous) of data points that are far apart from those data points that end up in a different cluster.**

- Our Example: K-means Clustering

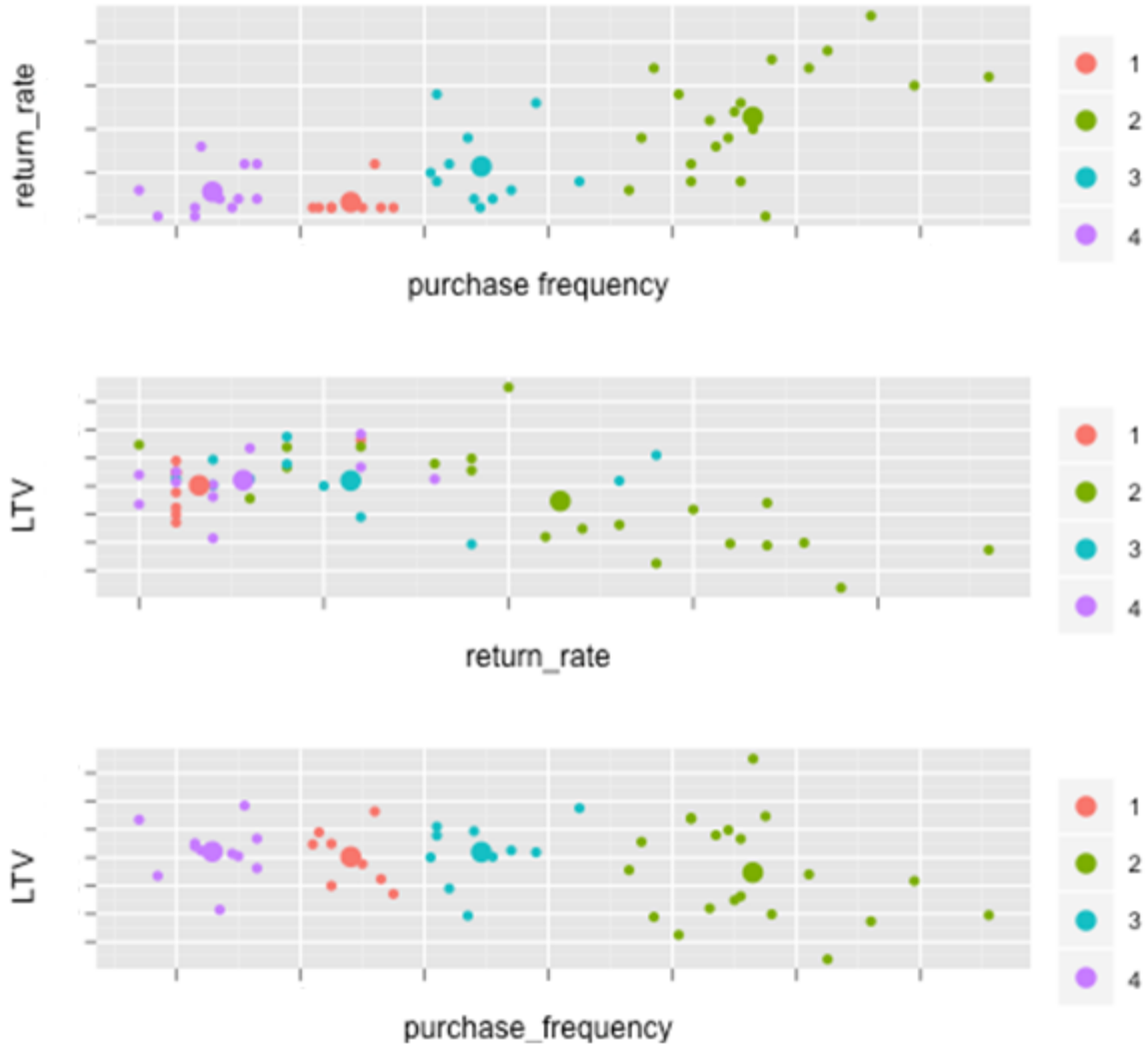
K-Means Clustering - What is it?

- Used for clustering numerical data, usually a set of measurements about objects of interest.
- Input: numerical. There must be a distance metric defined over the variable space.
  - Euclidian distance
- Output: The centers of each discovered cluster, and the assignment of each input to a cluster.
  - Centroid

Use Cases

- Often an exploratory technique:
  - Discover structure in the data
  - Summarize the properties of each cluster
- Sometimes an introduction to classification:
  - "Discovering the classes“
- Examples
  - The height, weight and average lifespan of animals
  - Household income, yearly purchase amount in dollars, number of household members of customer households
  - Patient record with measures of BMI, HBA1C, HDL

Use-Case Example – On-line Retailer



LTV – Lifetime Customer Value

Use-Case Example – On-line Retailer

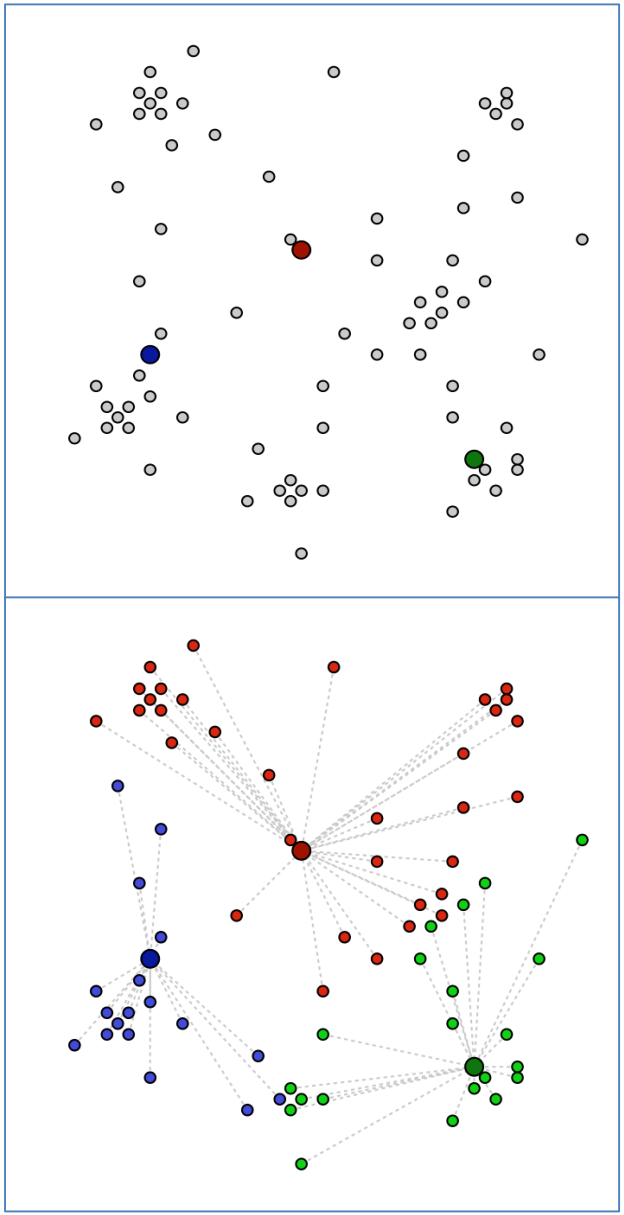
- Some of questions that a Data Scientist can address with the initial analysis with k-means clustering.
- Why is a certain group is ideal?
- What are the people in these different groups buying?
- Is that affecting LTV?
- Can we raise the LTV of our frequent customers, perhaps by lowering the cost of returns, or by somehow discouraging customers who return goods too frequently?
- Can we encourage customers to visit more (without lowering their LTV?)
- Are more frequent customers more valuable?

The Algorithm

1. Choose *K*; then select K random "centroids"

   In our example, K=3

2. Assign records to the cluster with the closest centroid



The Algorithm (Continued)

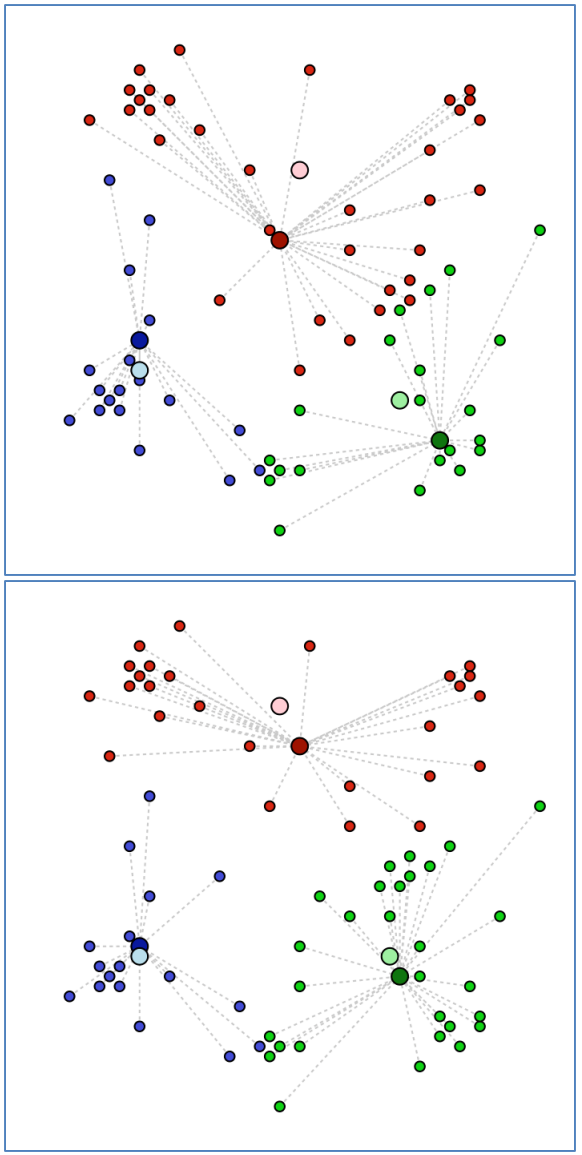
3. Recalculate the resulting centroids

    Centroid: the mean value of all the records in the cluster

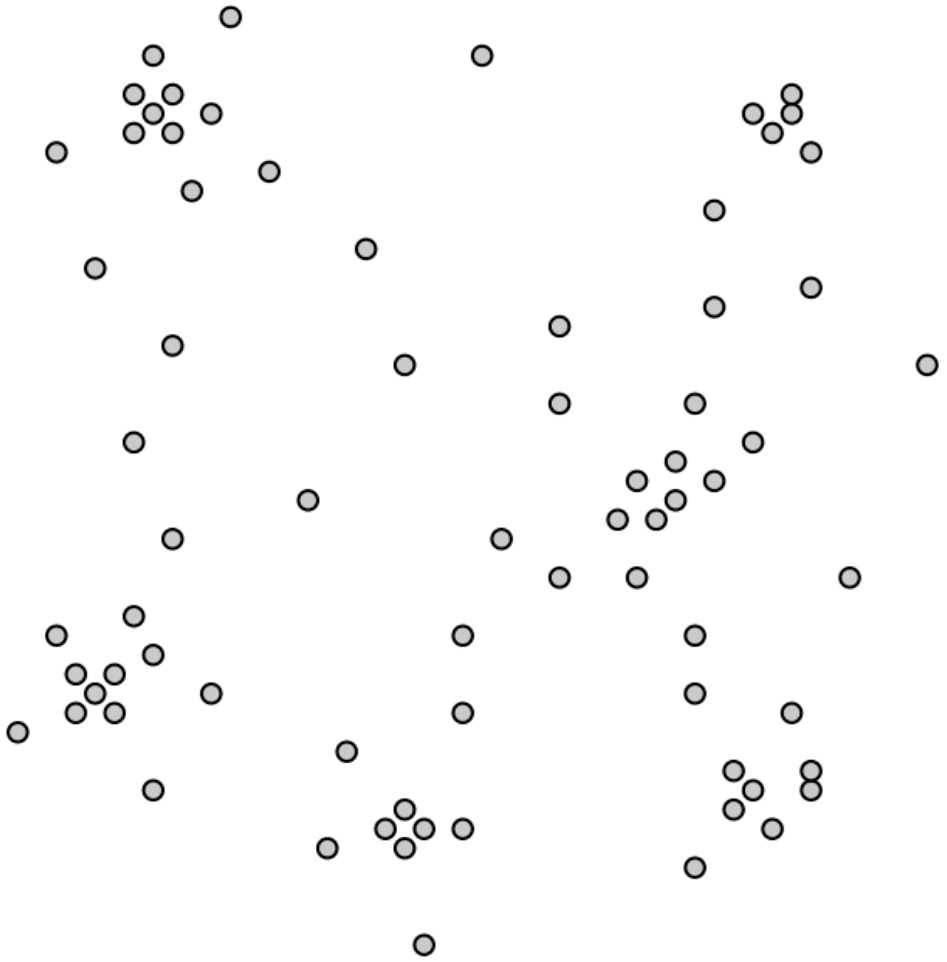
4. Repeat steps 2 & 3 until record assignments no longer change

Model Output:

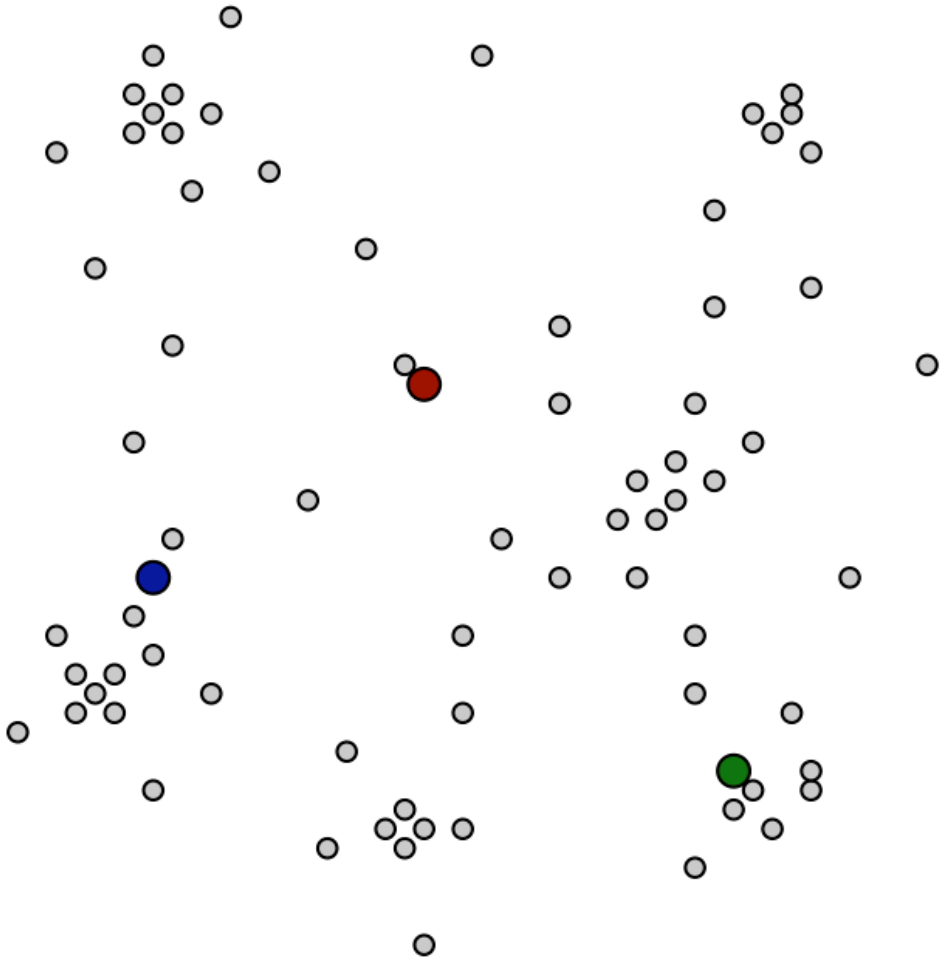
- The final cluster centers
- The final cluster assignments of the training data



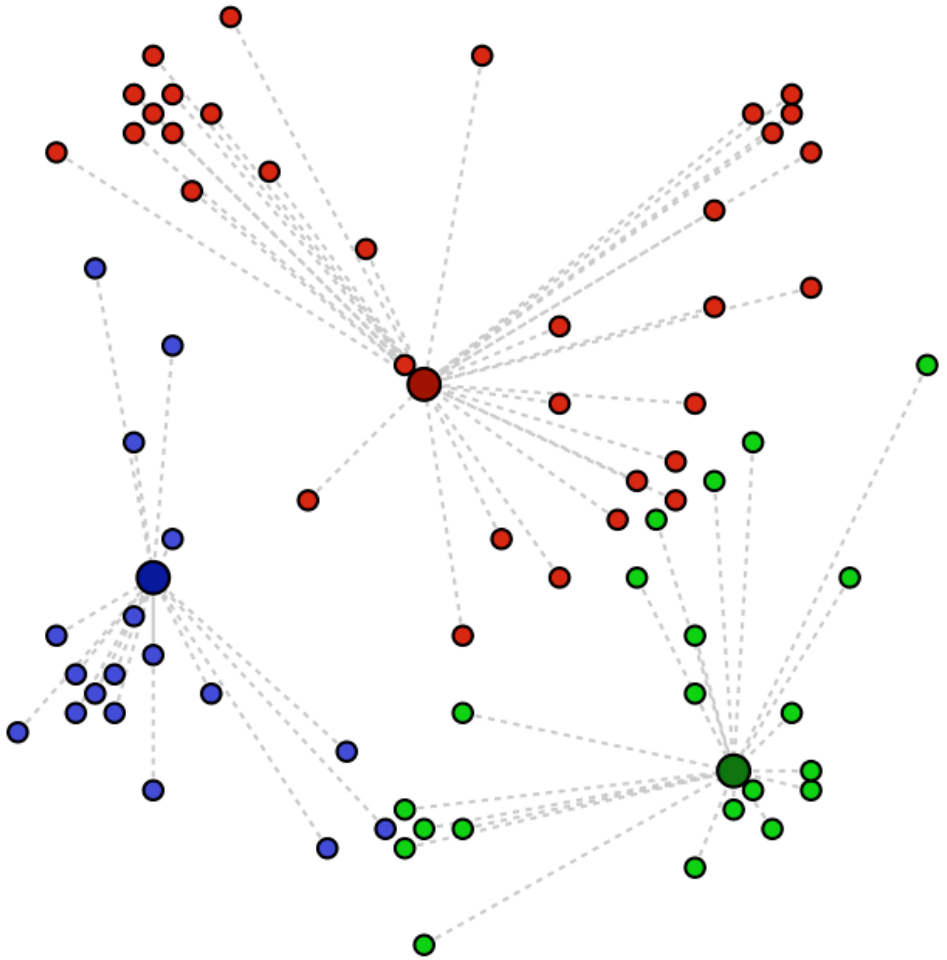
The Algorithm



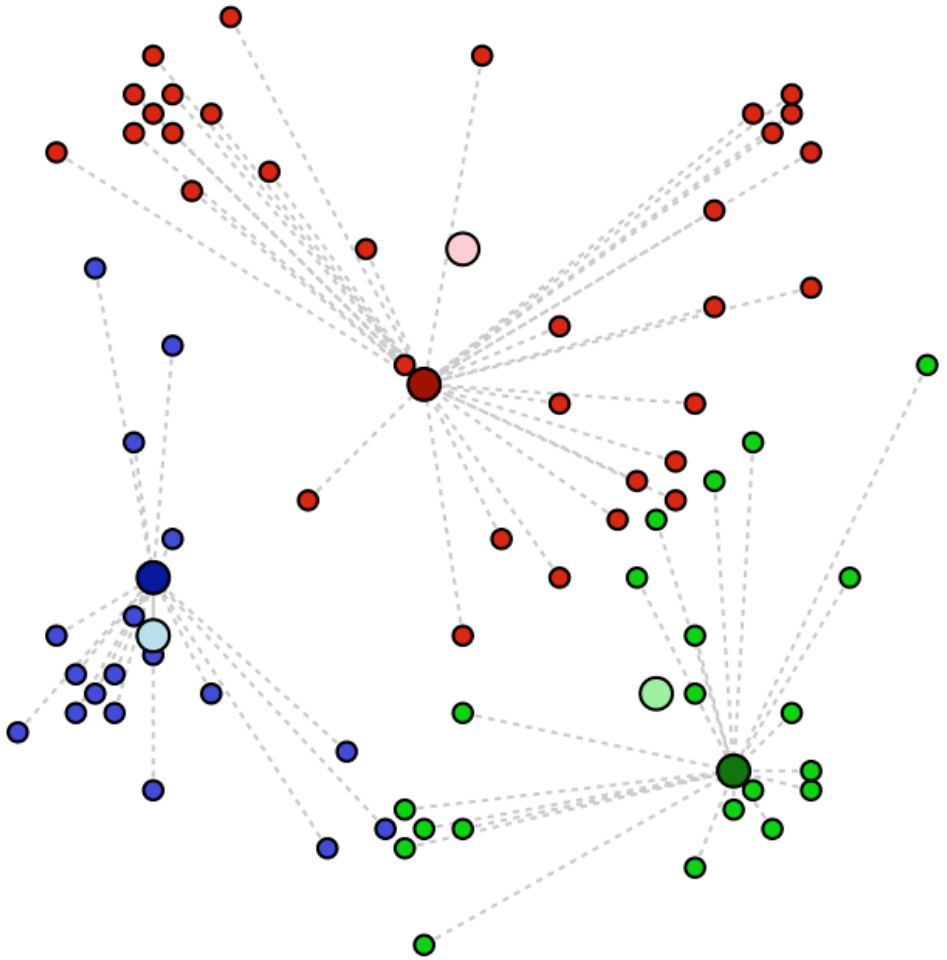
The Algorithm



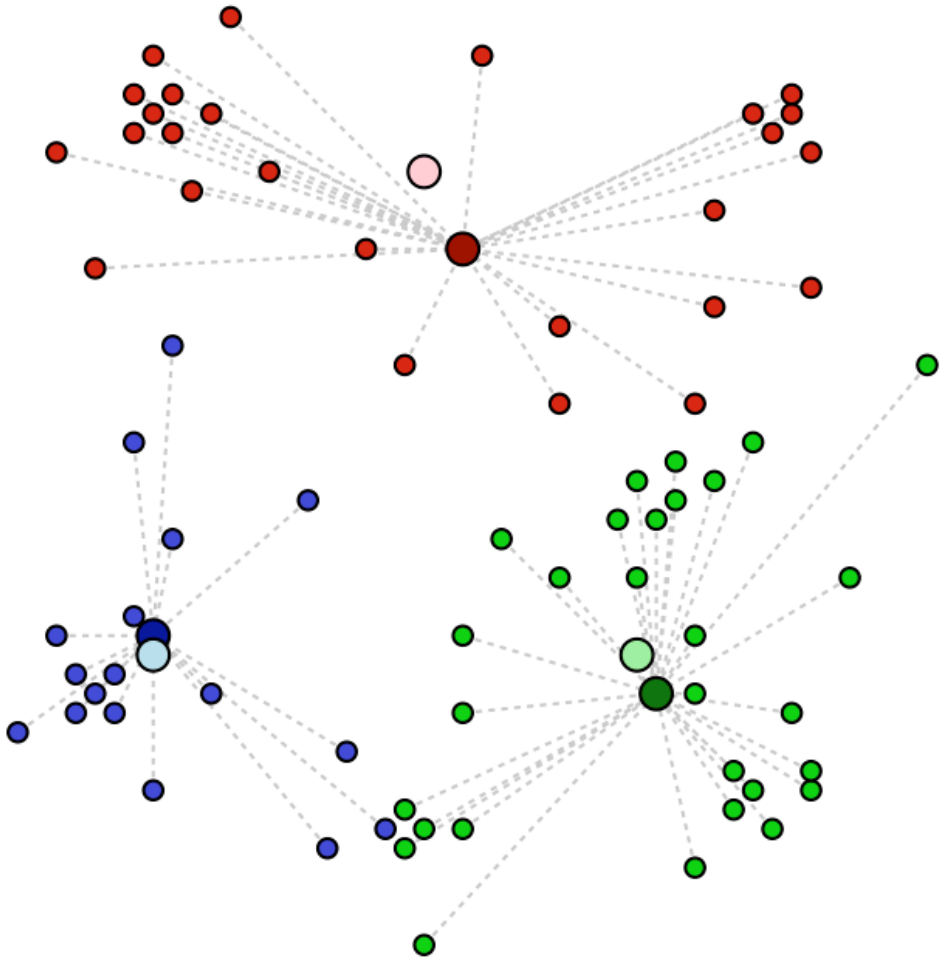
The Algorithm



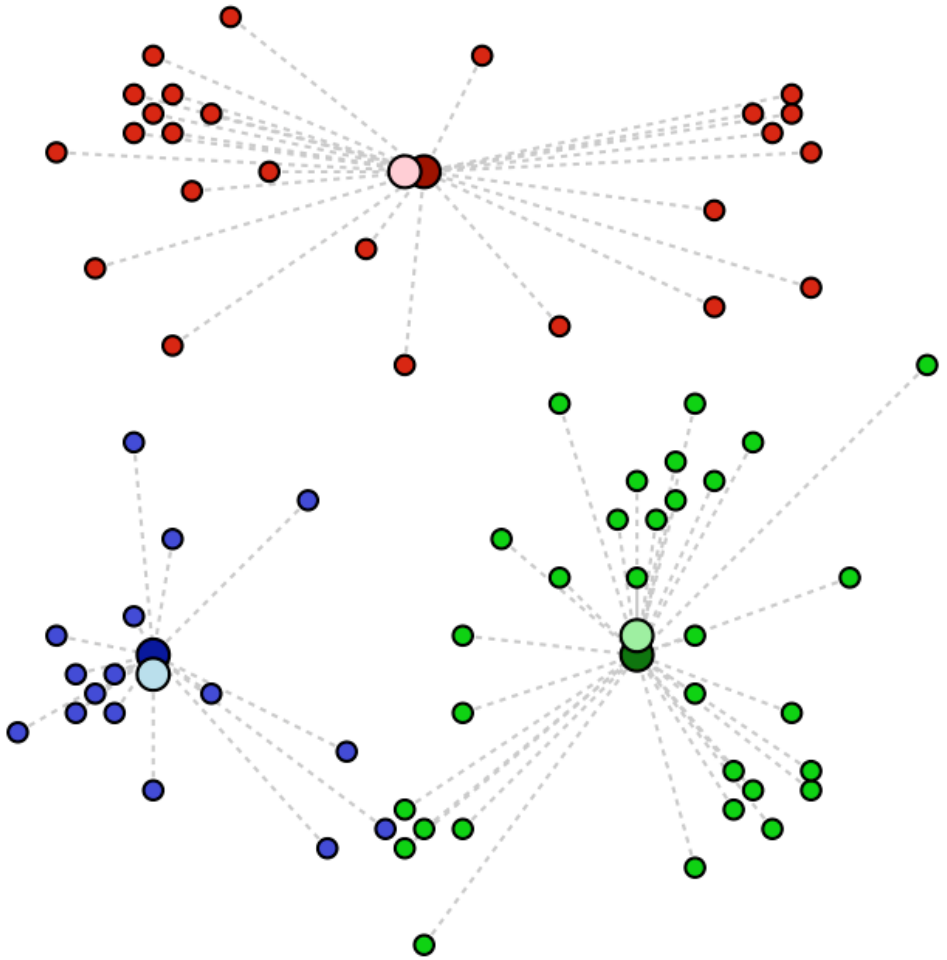
The Algorithm



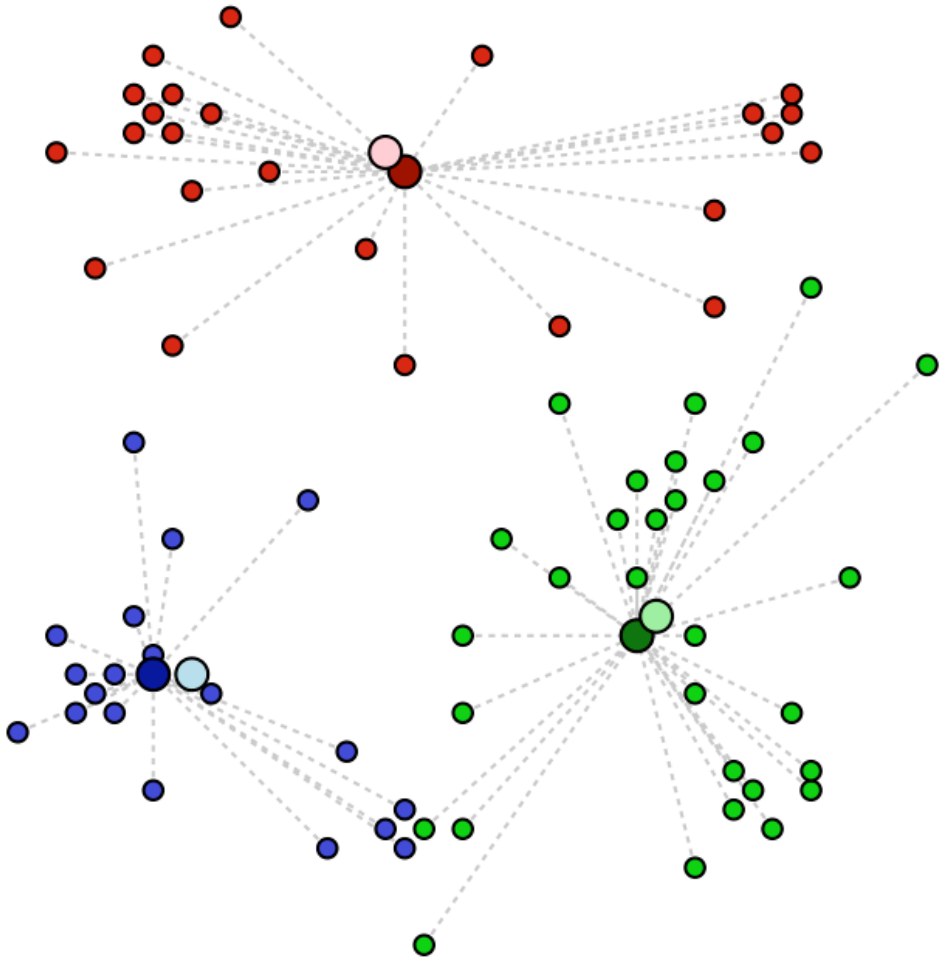
The Algorithm



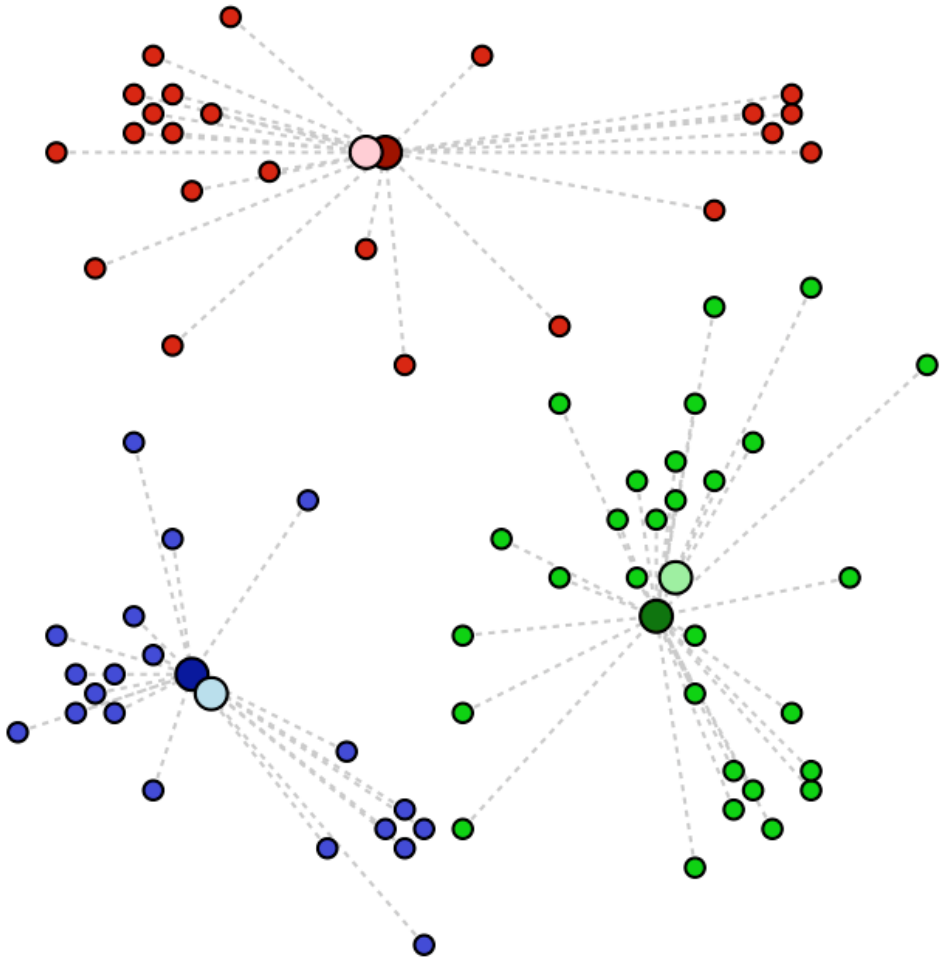
The Algorithm



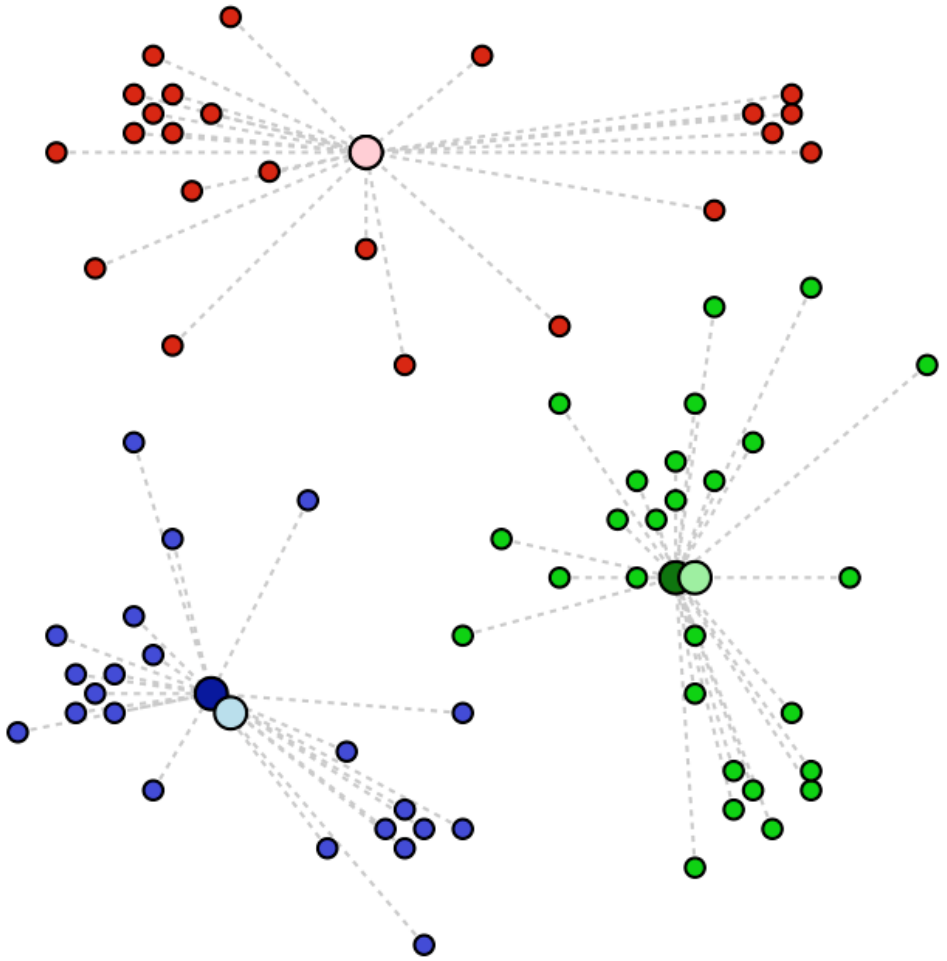
The Algorithm



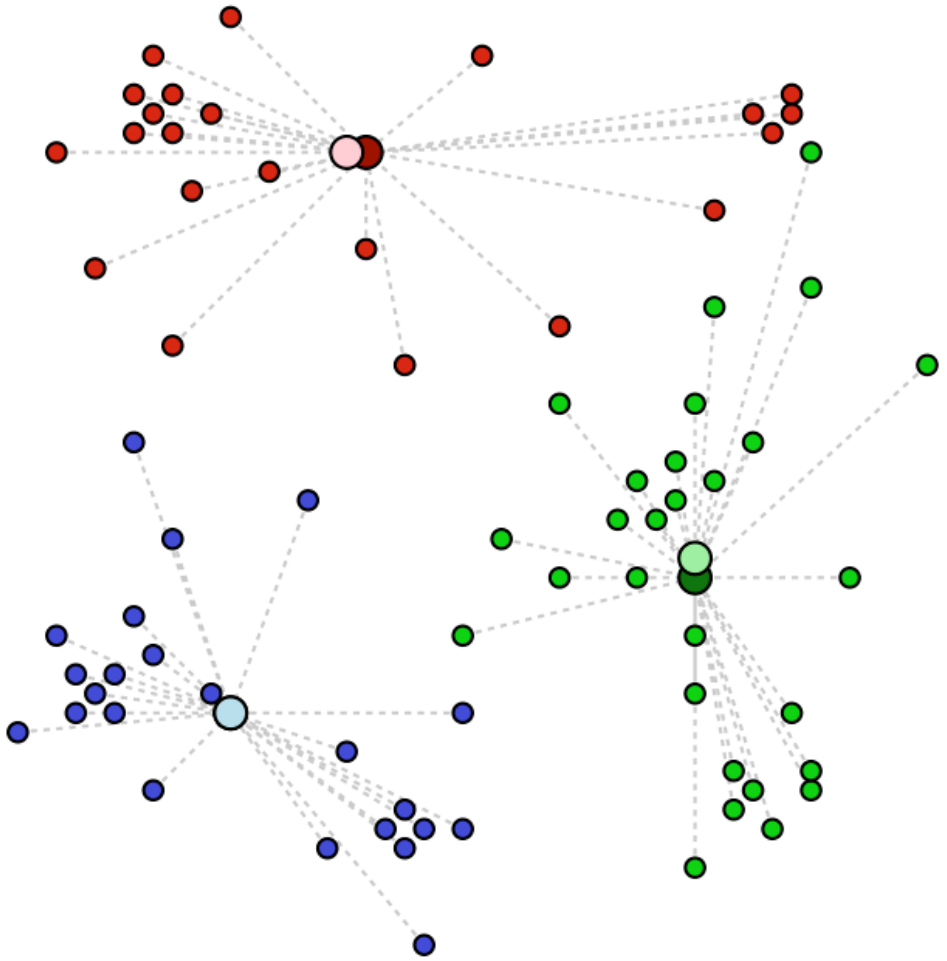
The Algorithm



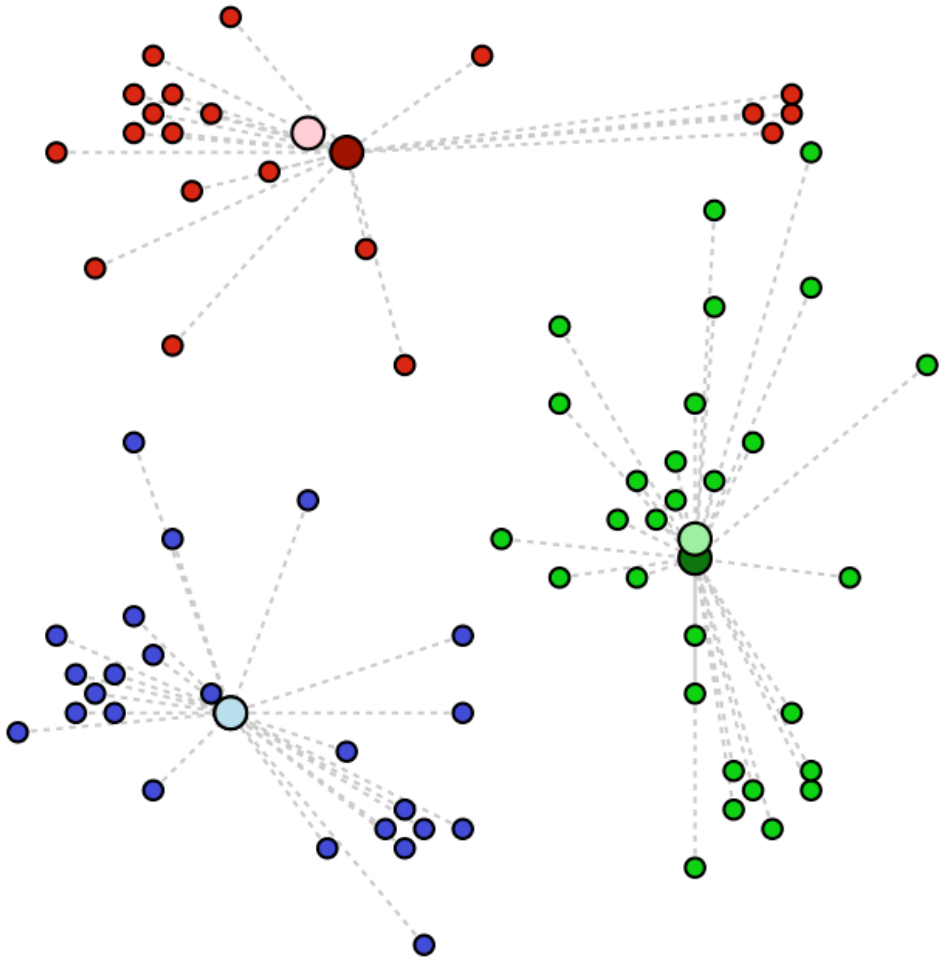
The Algorithm



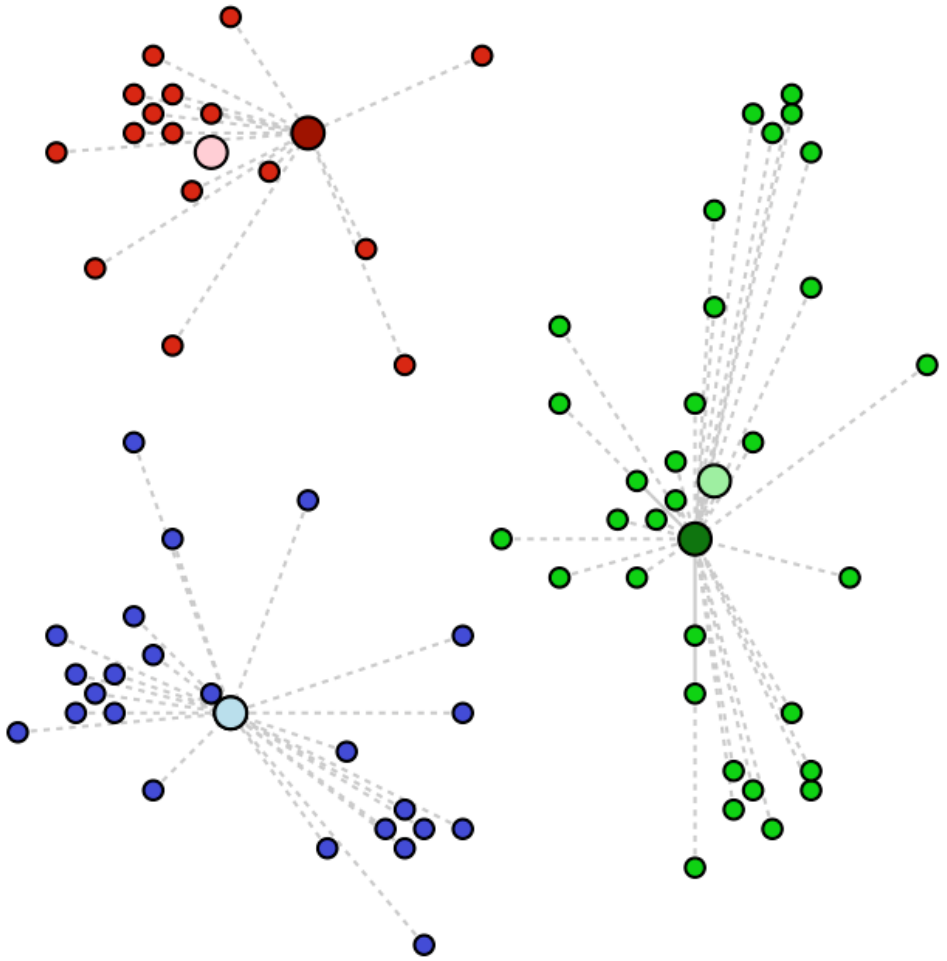
The Algorithm



The Algorithm



The Algorithm



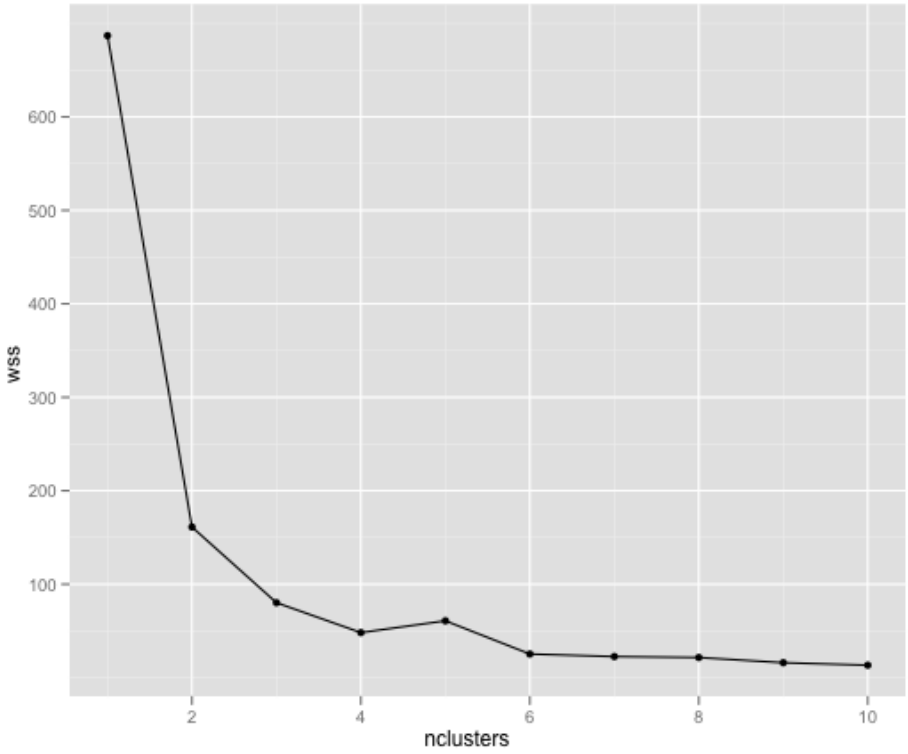
Picking K

Heuristic: find the "elbow" of the within-sum-of-squares (wss) plot as a function of K.

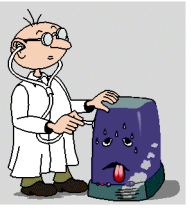
$$\text{wss} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - c_i|^2$$

K: # of clusters
n_i: # points in i^{th} cluster
c_i: centroid of i^{th} cluster
x_{ij}: jth point of i^{th} cluster

"Elbows" at k=2,4,6



Diagnostics – Evaluating the Model



- Do the clusters look separated in at least some of the plots when you do pair-wise plots of the clusters?
  - Pair-wise plots can be used when there are not many variables
- Do you have any clusters with few data points?
  - Try decreasing the value of K
- Are there splits on variables that you would expect, but don't see?
  - Try increasing the value K
- Do any of the centroids seem too close to each other?
  - Try decreasing the value of K

K-Means Clustering - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Easy to implement	Doesn't handle categorical variables
Easy to assign new data to existing clusters Which is the nearest cluster center?	Sensitive to initialization (first guess)
Concise output Coordinates the K cluster centers	Variables should all be measured on similar or compatible scales Not scale-invariant!
	K (the number of clusters) must be known or decided a priori Wrong guess: possibly poor results
	Tends to produce "round" equi-sized clusters. Not always desirable

Check Your Knowledge

Your Thoughts?

1. Why do we consider K-means clustering as a unsupervised machine learning algorithm?
2. How do you use “pair-wise” plots to evaluate the effectiveness of the clustering?
3. Detail the four steps in the K-means clustering algorithm.
4. How do we use WSS to pick the value of K?
5. What is the most common measure of distance used with K-means clustering algorithms?
6. The attributes of a data set are “purchase decision (Yes/No), Gender (M/F), income group (<10K, 10-50K, >50K). Can you use K-means to cluster this data set?

Module 4: Advanced Analytics – Theory and Methods

part 1: K-means Clustering - Summary

During this lesson the following topics were covered:

- Clustering – Unsupervised learning method
- What is K-means clustering
- Use cases with K-means clustering
- The K-means clustering algorithm
- Determining the optimum value for K
- Diagnostics to evaluate the effectiveness of K-means clustering
- Reasons to Choose (+) and Cautions (-) of K-means clustering

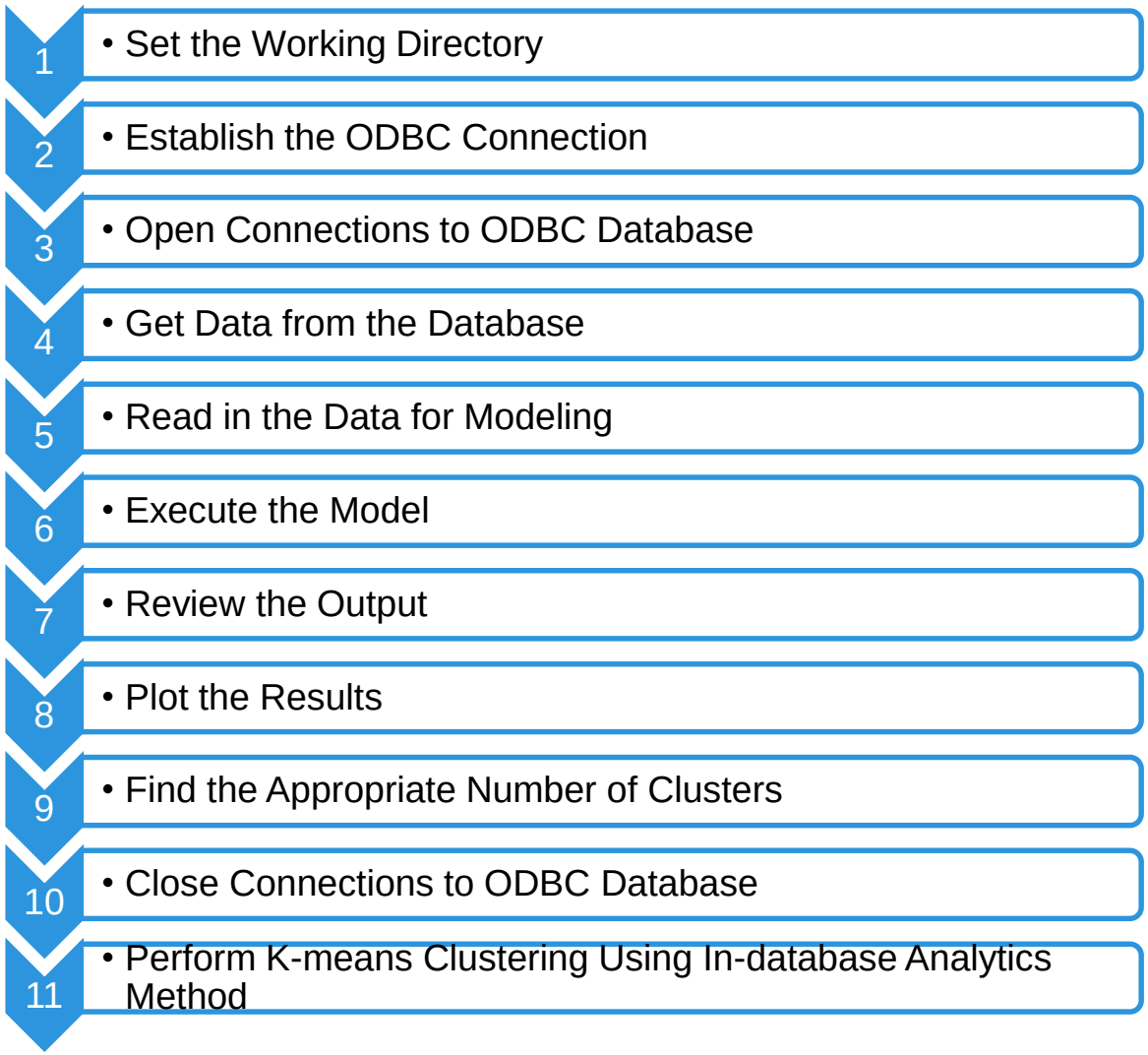
Lab Exercise 4: K-means Clustering

- This Lab is designed to investigate and practice K-means Clustering.

  After completing the tasks in this lab you should be able to:

  - Use R functions to create K-means Clustering models
  - Use ODBC connection to the database and execute SQL statements and read database tables in an R environment
  - Visualize the effectiveness of the K-means Clustering algorithm using graphic capabilities in R
  - Use MADlib function for K-means Clustering

Lab Exercise 4: K-means Clustering - Workflow



Thanks